Mitigating Perception Bias: A Training-Free Approach to Enhance LMM for Image Quality Assessment

Baoliang Chen[1*], Siyi Pan[1*], Dongxu Wu[1], Liang Xie[2†], Xiangjie Sui[3], Lingyu Zhu[4], Hanwei Zhu[5]

[1]School of Computer Science, South China Normal University, China
[2]School of Computer Science and Technology, Guangdong University of Technology, China
[3]Faculty of Data Science, City University of Macau, China
[4]School of Computer Science, City University of Hong Kong, China
[5]School of Computer Science and Engineering, Nanyang Technological University, Singapore
sypan@m.scnu.edu.cn, blchen@scnu.edu.cn

Abstract

Despite the impressive performance of large multimodal models (LMMs) in high-level visual tasks, their capacity for image quality assessment (IQA) remains limited. One main reason is that LMMs are primarily trained for high-level tasks (*e.g.*, image captioning), emphasizing unified image semantics extraction under varied quality. Such semantic-aware yet quality-insensitive perception bias inevitably leads to a heavy reliance on image semantics when those LMMs are forced for quality rating. In this paper, instead of retraining or tuning an LMM costly, we propose a training-free debiasing framework, in which the image quality prediction is rectified by mitigating the bias caused by image semantics. Specifically, we first explore several semantic-preserving distortions that can significantly degrade image quality while maintaining identifiable semantics. By applying these specific distortions to the query/test images, we ensure that the degraded images are recognized as poor quality while their semantics mainly remain. During quality inference, both a query image and its corresponding degraded version are fed to the LMM along with a prompt indicating that the query image quality should be inferred under the condition that the degraded one is deemed poor quality. This prior condition effectively aligns the LMM's quality perception, as all degraded images are consistently rated as poor quality, regardless of their semantic variance. Finally, the quality scores of the query image inferred under different prior conditions (degraded versions) are aggregated using a conditional probability model. Extensive experiments on various IQA datasets show that our debiasing framework could consistently enhance the LMM performance.

Code — https://barrypan12138.github.io/Q-Debias/

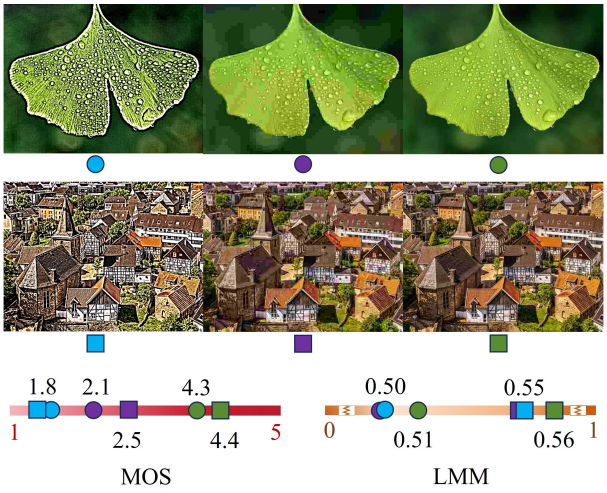

Figure 1. Illustration of perception bias in Large Multimodal Model (LMM) during quality assessment. Image quality ratings from the LMM (mPLUG-Owl3 (Ye et al. 2024)) were obtained using the Q-Bench testing framework (Wu et al. 2023a). The LMM consistently assigns higher quality ratings to images in the second row compared to the first, despite both sets exhibiting similar quality distributions as measured by Mean Opinion Scores (MOSs). This discrepancy suggests that the LMM relies more on image semantics than on low-level image clues for quality assessment.

Introduction

No-Reference Image Quality Assessment (NR-IQA) models aim to measure image quality in alignment with human perception without any reference, playing a fundamental role across various computer vision tasks (Wang and Bovik 2006). In the past decades, significant progress has been witnessed in NR-IQA, including traditional hand-crafted feature-based models (Chen et al. 2025) and deep learning-based models (Chen et al. 2021). Despite the advancement, current models still grapple with limited generalization capabilities on unseen scenes and distortions due to the significant distribution shifts between training and test sets. Recently, the emergence of Large Multimodal Models (LMMs) (OpenAI 2023) has demonstrated impressive generalization abilities across various vision-language tasks, such as classification (Wang et al. 2023), image captioning (Yu et al. 2023), and visual question answering (Shang et al. 2024).

*These authors contributed equally.
†Corresponding author.

However, the focus on high-level visual tasks usually limits their effectiveness in low-level tasks, such as IQA (Wu et al. 2023a; Sun et al. 2024). Although pioneering researchers have attempted to fine-tune or retrain LMMs for improved IQA accuracy (Wu et al. 2023b; Chen et al. 2024; Zhu et al. 2024), the laborious dataset construction and costly model training usually render this approach inefficient. Moreover, tuning LMMs specifically for IQA also introduces the risk of catastrophic forgetting (Luo et al. 2023), compromising the retention of general knowledge and ultimately degrading the models' capability on other tasks.

A powerful strategy to both retain the LMM's strengths across tasks and enhance its IQA performance is to unlock its vast general knowledge through well-crafted prompts, encouraging the LMM to respond accurately to quality rating requests. However, despite this potential, LMMs, driven by training objectives that emphasize semantic extraction over quality, usually default to ***interpreting image quality heavily relying on image semantics***. As shown in Fig. 1, two sets of distorted images with different semantics are presented to an LMM for quality rating. The results indicate that the LMM consistently prefers the quality of images with the second set of semantics, despite both sets having similar quality (Mean Opinion Score (MOS)) distributions. The case reveals that the LMM intrinsically bases its quality rating on image semantics rather than on quality-related clues (*e.g.*, blur). Motivated by this observation, we introduce an innovative, training-free approach to mitigate the perception bias inherent in LMMs. Our enhancement strategy consists of two main steps: 1) ***bias exposure***, and 2) ***bias mitigation***.

In the bias exposure step , we assume the bias exists consistently across images sharing the same semantics. Based on this assumption, we can expose the perception bias of a query/test image by measuring ***how much the LMM resists labeling it as high quality when its quality is severely degraded but the semantics are preserved.*** To achieve this, we explore several specific distortions that drastically corrupt the image quality while preserving its semantics to some extent. For a query image, we then impose these distortions on the query image and obtain its degraded versions. Herein, those degraded images should be deemed as poor quality. However, the LMM may not always agree with the fact and the disagreement leads to the bias exposure.

In the bias mitigation step, we address the exposed bias using an instructive conditional prompt. Specifically, during inference, we provide both the query image and its degraded version to the LMM, along with a prompt indicating that ***the quality of the query image should be assessed under the condition that the degraded counterpart is rated as poor quality***. By forcing the LMM to rate the quality of the degraded images appropriately, the bias mitigation in turn refines the LMM's quality prediction for the query image. Finally, the quality predictions under different distortions are aggregated through a conditional probability model, further improving the prediction accuracy. Before delving into detail, we highlight our main contributions as follows:

- **Investigation of Perception Bias in LMM for IQA.** We explore the perception bias inherent in LMM when used for IQA. Our training-free approach, which requires no task-specific fine-tuning, highlights a new pathway for leveraging pre-trained LMM on unseen tasks.
- **Conditional Prompt for Bias Mitigation.** We introduce a simple yet effective conditional prompt to mitigate the semantic bias in quality assessment. The prompt encourages the LMM to rate the image quality by aligning the quality prediction of synthetically degraded images, effectively reducing the bias caused by the semantics varies. Additionally, a confidence-based quality aggregation model is designed, further enhancing the prediction accuracy.
- **Comprehensive Evaluation on Diverse Datasets and Distortions.** We extensively evaluated our method on both natural and AI-generated images and the superior performance underscores the high effectiveness of our bias mitigation strategy. In addition, consistent improvements across multiple LMMs also demonstrate the strong generalization of our method, highlighting its potential to successfully extend to future LMMs.

Methodology

Preliminary

Given a query image x, the typical prompt for the LMM in IQA is exemplified as follows:
#User: Rate the quality of the image. Good or poor? (Question) [IMAGE_TOKEN](Image)
#Assistant: The quality of the image is [SCORE_TOKEN].
Based on the predicted logits of 'good' token (x^{gd}) and 'poor' token (x^{pr}) on the position of [SCORE_TOKEN], the image quality score y can be estimated by a **SoftMax** function:

$$p(y \mid x) = \frac{e^{x^{\text{gd}}}}{e^{x^{\text{gd}}} + e^{x^{\text{pr}}}}. \tag{1}$$

However, the semantic bias inherent in the LMM usually results in unreliable quality estimation, as the inference heavily relies on image semantics. To account for this, we adopt a conditional probability model to mitigate the bias, which can be formulated as follows,

$$p(y \mid x) = \mathrm{E}_{x' \mid x} p\left(y \mid x, x'\right) p\left(x' \mid x\right), \tag{2}$$

where x' is a "conditional image" of x, whose quality has been severely degraded while retaining similar semantics to x. $p\left(x' \mid x\right)$ represents the probability distribution of the potential degradation results. During inference, both the conditional image and the query image are fed to the LMM with a prompt instructing the LMM to rate the quality of the query image, under the condition that the conditional image is considered of poor quality. Our design philosophy is to guide the LMM toward confidently and accurately classifying the degraded images as poor quality, reducing its high reliance on image semantics in quality inference. This bias mitigation can, in turn, be propagated to the query image quality inference, assuming that the bias is consistently present in images with similar semantics but varying distortions.

Framework of Perception Bias Mitigation

Guided by the model constructed in Eqn. (2), we design our framework mainly comprises two components: 1) Bias

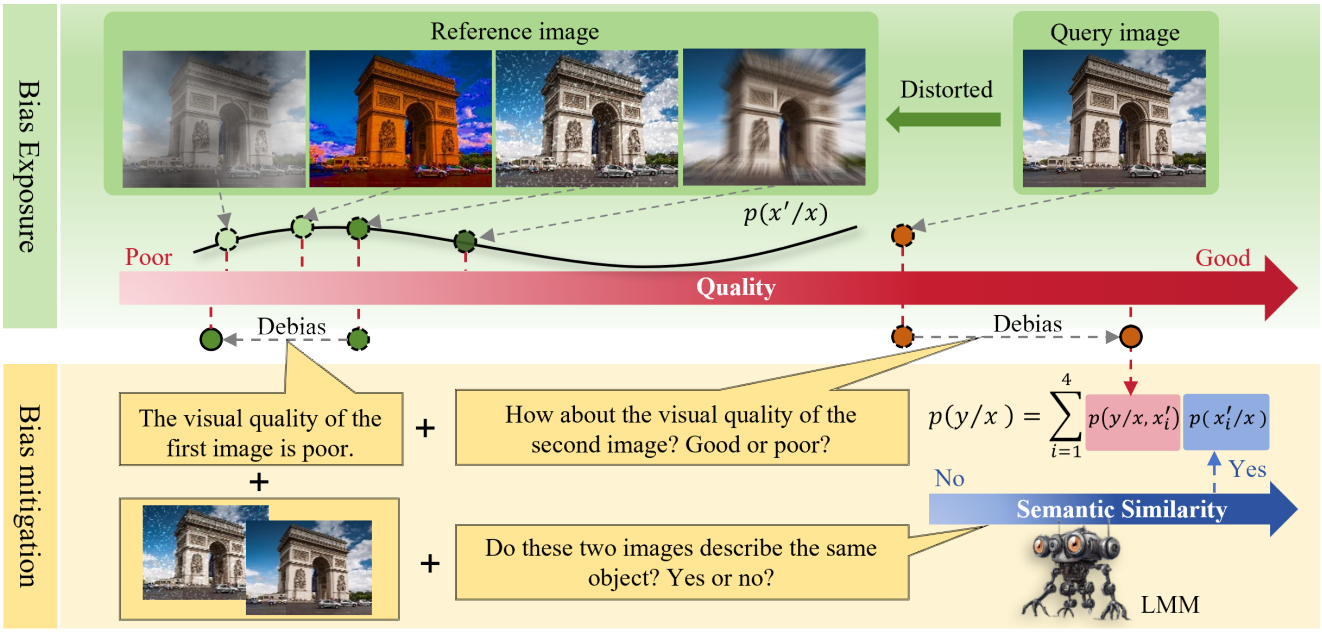

Figure 2. The framework of our perception bias mitigation scheme. It mainly consists of two components: 1) Bias Exposure: Specific distortions are imposed on the query image to significantly degrade the query image quality while preserving its semantics. The disagreement that the LMM rates those distorted images as poor quality exposes the perception bias inherent in the LMM. 2) Bias Mitigation: Dedicated prompts are defined to mitigate the bias by forcing that the quality of the query image should be assessed under the condition that its degraded counterpart is rated as poor quality. The final quality is then estimated by a semantic similarity based aggregation.

Exposure. Specific distortions that significantly degrade image quality while preserving semantics are explored and imposed on the query image to construct different conditional images $p\left(x' \mid x\right)$. 2) Bias Mitigation. A dedicated prompt is designed to estimate $p\left(y \mid x, x'\right)$ across different conditional images and obtain the final quality by Eqn. (2). Our framework is illustrated in Fig. 2 and each component is detailed as follows.

Bias Exposure Given a query image, we examine four typical distortions—zoom blur, spatter noise, saturation enlargement, and fog corruption—to effectively degrade the image quality while preserving its semantic content. Specifically:

Zoom Blur. The zoom blur distortion usually occurs when a camera moves toward an object rapidly, which can be simulated by

$$x_1'(u,v) = \frac{1}{n}\sum_{i=1}^{n} x_{z_i}(u,v), \tag{3}$$

where x_{z_i} means the zoom result of the query image x by a factor z_i and a total of n factors are adopted. $x_0'(u,v)$ means the zoom blur results at position (u,v).

Spatter Noise. We use spatter noise to mimic the distortion caused by an unclean camera lens due to bad weather conditions such as rain, mud, or dust, which can be generated by

$$x_2'(u,v) = x(u,v)\cdot(1-M(u,v)) + C(u,v)\cdot M(u,v), \tag{4}$$

where $M(u,v)$ is the spatter mask indicating regions that are affected or unaffected, and $C(u,v)$ represents the specific color distribution adapted for different splatter types. Herein, we use the implementation in (Hendrycks and Dietterich 2019) to generate $M(u,v)$ and $C(u,v)$.

Saturation Enlargement. The saturation distortion modifies the saturation channel of an image in the HSV color space based on the severity parameter c_0, which can be defined by

$$x_3' = f_{hsv2rgb}(x_h, x_s', x_v), \tag{5}$$

with

$$x_s' = \text{clip}\left(x_s \cdot c, 0, 1\right), \tag{6}$$

where x_h, x_s, and x_v represent the hue, saturation, and value components of x in HSV space, respectively. x_s' denotes the distorted saturation. The function $f_{hsv2rgb}(\cdot)$ converts the image from HSV to RGB space, while $\text{clip}(\cdot)$ ensures that the distorted results are clipped to a valid range.

Fog Corruption. We simulate a foggy environment by applying a haze effect to the query image as follows:

$$x_4' = \text{clip}\left(x + k \cdot x^F, 0, 1\right), \tag{7}$$

where the x^F is the fog pattern generated by a diamond-square algorithm (Fournier, Fussell, and Carpenter 1982;

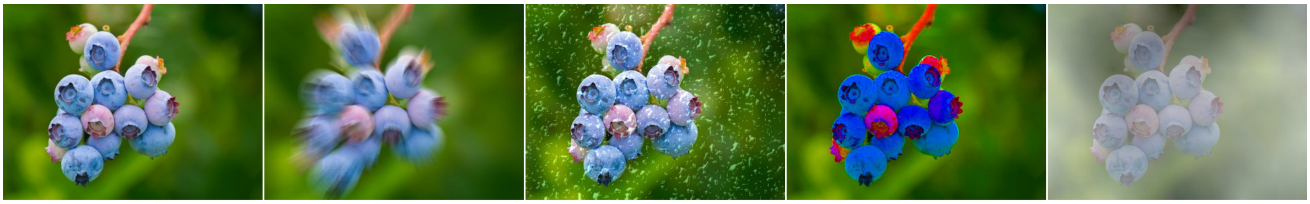

Figure 3. Illustration of the four distortion types which could degrade image quality significantly while largely preserving its semantics.

Hendrycks and Dietterich 2019) based on x. k is the hyper-parameter of distortion severity. As shown in Fig. 3, the four types of distortion degrade the image quality significantly while the semantics are still recognizable, leading to an expected bia exposure when they are not deemed poor quality by the LMM.

Bias Mitigation Based on the generated conditional images for each query image, we then input the query image (x) and one of its counterparts (x_i') into the LMM, using a specific prompt to propagate the bias mitigation effect from the conditional image to the query image.

#User: The visual quality of the first image is poor. How about the visual quality of the second image? Good or poor? (Question) [IMAGE_TOKEN1, IMAGE_TOKEN2].(Image1, Image2)

#Assistant: The quality of the image is [SCORE_TOKEN].

Then, the conditional quality probability can be estimated as follows:

$$p(y \mid x, x_i') = \frac{e^{x^{\mathrm{gd}}}}{e^{x^{\mathrm{gd}}} + e^{x^{\mathrm{pr}}}}. \tag{8}$$

Finally, we aggregate the quality estimation across the four distortion types:

$$p(y \mid x) = \sum_{i=1}^{4} p\left(y \mid x, x_i'\right) p\left(x_i{}' \mid x\right), \tag{9}$$

where $p\left(x_i' \mid x\right)$ is the probability that the distorted image is adopted as the condition. We leverage the semantic similarity between x and x' to estimate this probability, based on the assumption that the more semantic information maintained, the more confidently the image can be considered as a condition. We achieve the semantic similarity estimation by feeding another prompt to the LMM as follows,

#User: Do these two images describe the same object? Yes or no? (Question)

#Assistant: [SCORE_TOKEN]. This yields

$$p\left(x_i{}' \mid x\right) = \frac{e^{w_i}}{\sum_{i=1}^{4} e^{w_i}}, \tag{10}$$

with

$$w_i = \frac{e^{x_i^{\mathrm{yes}}}}{e^{x_i^{\mathrm{yes}}} + e^{x_i^{\mathrm{no}}}}. \tag{11}$$

Experiments

Experimental Settings

Datasets. We evaluate our method on five publicly available datasets: LIVE Challenge (Ghadiyaram and Bovik 2015), KonIQ-10k (Hosu et al. 2020), AGIQA-3k (Li et al. 2023), KADID-10k (Lin, Hosu, and Saupe 2019), and SPAQ (Fang et al. 2020). The KonIQ-10k, SPAQ, and LIVE Challenge datasets are in-the-wild image collections, featuring authentic distortions. KonIQ-10k and SPAQ datasets each contain over 10,000 images and the SPAQ dataset is specifically designed to assess the quality of images captured by smartphones. The KADID-10k dataset comprises 10,125 images with various systematic distortions. The AGIQA-3k dataset includes 2,900 images focused on AI-generated image quality assessment.

Comparison Methods. We denote our method as "Q-Debias" and compare its performance against both training-free (opinion-unaware) and training-based methods across multiple datasets. The training-free methods include BLINDS-II (Moorthy and Bovik 2010), BRISQUE (Mittal, Moorthy, and Bovik 2012), NIQE (Mittal, Soundararajan, and Bovik 2012), NPQI (Liu et al. 2020), ContentSep (Babu, Kannan, and Soundararajan 2023), CLIP-IQA (Wang, Chan, and Loy 2023), MDFS (Ni et al. 2024), and Q-Bench (Wu et al. 2023a). In particular, Q-Bench refers to using the same LMM (mPLUG-Owl3) with our method, while applying the query prompt from Q-Bench. For the training-based methods, we adopt models trained on the large-scale KonIQ-10k dataset for comparison, including ARNIQA (Agnolucci et al. 2024), TReS (Golestaneh, Dadsetan, and Kitani 2022), and MUSIQ (Ke et al. 2021). We list their performance in cross-dataset settings to verify their generalization capability. The Pearson Linear Correlation Coefficient (PLCC) and the Spearman Rank-Order Correlation Coefficient (SRCC) are used as metrics to assess the linearity and monotonicity of our quality predictions.

Implementation Details. We select mPLUG-Owl3 as our LMM due to its superior performance on image processing tasks. We set $z_i \in \{1.00, 1.01, 1.02, \ldots, 1.10\}$, $n = 11$ in Eqn. (3), and $c = 2.0$ in Eqn. (6). We set $k = 2.5$ in Eqn. (7).

Comparison with NR-IQA Models

Prediction Accuracy. As shown in Table **??**, compared with existing training-free methods, our model Q-Debias con-

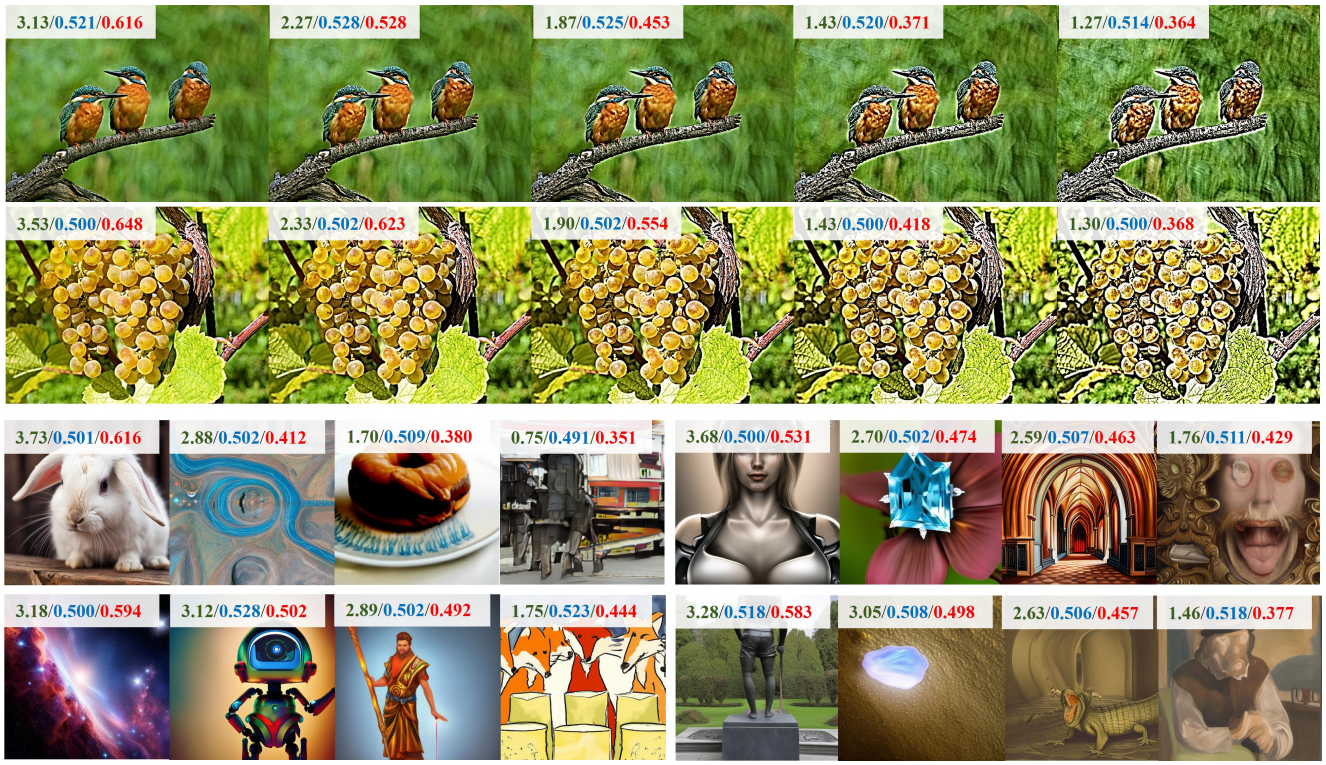

Figure 4. Visualization of image quality prediction results. In each subfigure, the top-left label shows numbers in green, blue and red, representing the MOS, the LMM prediction result with the prompt in Q-Bench and our result, respectively.

sistently achieves the best performance across the first five IQA datasets. In particular, most hand-crafted feature-based models (*e.g.*, NIQE) experience a significant performance drop on the KADID-10k dataset due to the diverse distortion types involved. However, our method still outperforms these models by a large margin, demonstrating its high effectiveness. Compared to the vanilla prompt used in Q-Bench, our method while utilizing the same base LMM (mPLUG-Owl3), achieves performance gains across all five datasets.

In comparison to training-free methods, training-based models generally deliver superior results, benefiting from the quality assessment knowledge learned from large-scale datasets. However, due to the fact that KonIQ-10k only contains authentic distortions, these models often underperform on unseen distortions when tested on datasets involving unseen distortions (e.g., TReS: 0.771 on LIVE Challenge vs 0.468 on KADID-10k, MUSIQ: 0.788 on LIVE Challenge vs 0.630 on AGIQA-3k), highlighting the overfitting dilemma during training. In contrast, our approach improves the LMM in a training-free manner, providing superior generalization capability across authentic, systemic, and AI-generated distortions.

Visualization. To verify the effectiveness of our method, we visualize our predicted results alongside the LMM predictions using the Q-Bench prompt on the KADID-10k dataset (first two rows) and the AIGC-3k dataset (second two rows).

As shown in Fig. 4, we can observe that: 1) Despite comparable distortions and closely aligned MOS distributions in the first two rows, the LMM without our debiasing enhancement consistently assigns higher quality ratings to images in the first row over the second. This observation underscores the model's high reliance on semantic content rather than low-level clues for quality assessment, revealing the presence of perceptual bias. 2) The bias varies by semantic content, affecting both natural and AI-generated images. In contrast, our debias strategy could effectively mitigate such bias, resulting in predictions that are more consistent with human ratings (*i.e.*, MOSs).

Generalization on other LMMs

In our method, we adopted the multimodal model mPLUG-Owl3 as our foundation model. To demonstrate the generalization capability of our enhancement strategy, we further validate it on another four LMMs inducing: Bak-LLaVA (Liu et al. 2024), Qwen-VL (Bai et al. 2023), LLaVA-OneVision (Li et al. 2024a), LLaVA-Interleave (Li et al. 2024b) and DeepSeek-VL2 (Wu et al. 2024). As shown in Table 2, we could observe a consistent average performance gains can be achieved. Notably, the improvements observed on the LIVE Challenge and AGIQA-3k datasets suggest that semantic bias is widespread across diverse content types and such bias can be mitigated by our method efficiently. The promising generalization capability highlights the transformative potential of our bias-mitigation strategy and opens up exciting new avenues for developing training-

Methods	Training-free? (Training set)	LIVE Challenge		KonIQ-10k		AGIQA-3k		KADID-10k		SPAQ	
		SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑
BLIINDS-II	✓	0.090	0.107	0.585	0.598	0.454	0.510	0.224	0.313	0.317	0.326
QAC	✓	0.226	0.284	0.340	0.291	-	-	0.239	0.309	0.440	0.450
BRISQUE	✓	0.561	0.598	0.705	0.707	0.493	0.533	0.330	0.370	0.484	0.481
NIQE	✓	0.463	0.491	0.551	0.488	0.528	0.520	0.379	0.389	0.703	0.671
ILNIQE	✓	0.439	0.503	0.505	0.496	0.594	0.623	0.540	0.534	0.696	0.637
NPQI	✓	0.475	0.490	0.613	0.614	0.658	0.714	0.391	0.340	0.600	0.616
ContentSep	✓	0.506	0.513	0.640	0.627	-	-	0.506	0.357	0.708	0.665
CLIP-IQA	✓	0.612	0.594	0.695	**0.727**	0.658	0.714	0.500	0.520	0.738	0.735
MDFS	✓	0.482	0.536	**0.733**	0.712	**0.672**	0.676	0.598	0.594	0.741	0.718
ARNIQA	✗ (KonIQ-10k)	0.670	0.715	-	-	0.621	0.694	**0.725**	**0.717**	0.576	0.577
TReS	✗ (KonIQ-10k)	0.771	0.805	-	-	0.652	**0.737**	0.468	0.492	0.418	0.417
MUSIQ	✗ (KonIQ-10k)	**0.788**	**0.824**	-	-	0.630	0.722	0.556	0.575	0.726	**0.738**
Q-Bench *	✓	0.721	0.677	0.672	0.573	0.596	0.469	0.315	0.267	**0.767**	0.650
Q-Debias (Ours)	✓	**0.794**	**0.818**	**0.838**	**0.863**	**0.717**	**0.753**	**0.700**	**0.753**	**0.867**	**0.826**
Improvement over Q-bench		**↑ 10.1%**	**↑ 20.8%**	**↑ 24.7%**	**↑ 50.6%**	**↑ 20.3%**	**↑ 60.6%**	**↑ 122.2%**	**↑ 167.0%**	**↑ 13.0%**	**↑ 27.1%**

* We use the same quality query prompt from Q-Bench and set the LMM as mPLUG-0wl3, consistent with our method.

Table 1. Performance comparison of our method, Q-Debias, against both training-free and training-based IQA models. The percentage indicates the improvement of our method over Q-Bench. The best two results are highlighted in boldface.

Models	LIVE Challenge				AGIQA-3k			
	Vanilla		Our debias		Vanilla		Our debias	
	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑
BakLLaVA	0.090	0.108	0.263	0.265	0.480	0.321	0.460	0.482
Qwen-VL	0.470	0.546	0.504	0.501	0.504	0.532	0.615	0.623
OneVision	0.379	0.654	0.631	0.649	0.581	0.781	0.707	0.806
Interleave	0.221	0.337	0.454	0.543	0.223	0.315	0.464	0.560
DeepSeek-VL2	0.800	0.851	0.822	0.860	0.606	0.655	0.759	0.778

Table 2. Performance improvement on other LMMs.

free enhancement methods to fully harness the potential of LMMs for unseen tasks.

Ablation Studies

Study of Conditional Images. In our method, distortion types are carefully selected to degrade image quality while preserving semantic content. Specifically, we explore five types of distortions for conditional image construction: **blur, noise, adverse weather conditions, brightness adjustment, and saturation modification**. For blur distortion, we consider zoom blur, motion blur, and Gaussian blur, while for noise distortion, we examine Gaussian noise and spatter noise. Additionally, we synthesize significant snow, frost, and fog distortions, as humans can still recognize objects in images captured under adverse weather conditions. The study results are summarized in Table 3.

From Table 3, we observe that all distortion types enhance AI-generated images in the AGIQA-3k dataset, whereas only a subset (*e.g.*, spatter noise and saturation modification) improves quality prediction for natural images in the LIVE Challenge dataset. This suggests that perception bias in AI-generated images is more pronounced, making performance gains more detectable. A potential explanation is that LMMs have been exposed to significantly fewer AI-generated images compared to natural images during training due to the vast historical disparity in dataset sizes. Moreover, different distortion types exhibit varying effectiveness in bias mitigation, underscoring the importance of careful distortion design, as bias levels are highly dependent on image semantics. To develop a generalized and effective bias mitigation strategy, we further explore potential combinations of the examined distortions. Given the exponential growth in possible combinations, we evaluate a four-distortion combination scheme, selecting the most effective distortion from each category. The results demonstrate that incorporating diverse distortion types yields the highest performance gains, highlighting their complementary roles in mitigating bias.

In our approach, distortions are applied directly to the query image to generate degraded versions while maintaining semantic consistency. To assess the necessity of **semantic consistency**, we further investigate the use of conditionally distorted images with mismatched semantics. Specifically, we construct an additional conditional image set by applying the four selected distortions to open-source images that do not share semantic content with those in the LIVE Challenge and AGIQA-3k datasets. For each quality inference, we randomly select four low-quality images from this set as conditional images. The results, presented in Exp. 14 of Table 3, show a significant performance drop when semantically inconsistent images are used, reinforcing the critical role of semantic alignment in bias mitigation. This finding highlights that bias is highly semantic-specific—leveraging conditionally degraded images that align with the query image semantics enables more accurate bias estimation and ultimately improves quality prediction.

Study of Instructive Prompt. In our method, the prompt serves an instructive role for the LMM, facilitating the propagation of bias mitigation from the conditional images to the query image. To verify its effectiveness, we compare

Exp. ID	Single Distortion		CS	LIVE Challenge		AGIQA-3k		Average	
				SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑
1	Blur	Zoom	✓	0.644	0.673	0.633	0.658	**0.639**	**0.666**
2		Motion	✓	0.497	0.515	0.552	0.546	0.525	0.531
3		Gaussian	✓	0.617	0.497	0.636	0.648	0.627	0.573
4	Noise	Gaussian	✓	0.677	0.646	0.686	0.720	0.682	0.683
5		Spatter	✓	0.762	0.799	0.713	0.768	**0.738**	**0.784**
6	Bad weather	Snow	✓	0.713	0.761	0.686	0.640	0.700	0.701
7		Frost	✓	0.632	0.705	0.633	0.573	0.633	0.639
8		Fog	✓	0.729	0.763	0.689	0.702	**0.709**	**0.733**
9	Brightness		✓	0.613	0.673	0.620	0.668	**0.617**	**0.671**
10	Saturation		✓	0.784	0.790	0.720	0.735	**0.752**	**0.763**

Multiple Distortions

Exp. ID	D1	D2	D3	D4	Sem	LIVE Challenge SRCC↑	LIVE Challenge PLCC↑	AGIQA-3k SRCC↑	AGIQA-3k PLCC↑	Average SRCC↑	Average PLCC↑
11	✗	✗	✗	✗	✗	0.721	0.677	0.596	0.469	0.659	0.573
12	✗	✓	✗	✓	✓	0.793	0.790	0.709	0.712	0.751	0.763
13	✗	✓	✓	✓	✓	0.793	0.773	0.714	0.702	0.753	0.738
14	✓	✓	✓	✓	✗	0.493	0.472	0.518	0.508	0.506	0.490
Q-Debias	✓	✓	✓	✓	✓	**0.794**	**0.818**	**0.717**	**0.753**	**0.756**	**0.786**

* CS: Semantic Consistency; Sem: Semantic. D1, D2, D3, and D4 represent Zoom, Spatter, Fog, and Saturation, respectively.

Table 3. Ablation study of different types of conditional images. The best results are highlighted in boldface.

Prompt	LIVE Challenge		KonIQ-10k		AGIQA-3k	
	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑
T1	0.784	0.762	0.805	0.816	0.703	0.682
T2	0.785	0.762	0.813	0.847	0.705	0.702
T3	0.741	0.730	0.811	0.845	0.672	0.686
Q-Debias	**0.794**	**0.818**	**0.838**	**0.863**	**0.717**	**0.753**

Table 4. Ablation study of instructive prompt and Aggregation Scheme.

our method against three prompt variants: **(T1)** Prompt Replacement: The entire prompt is replaced with "Rate the quality of the second image. Good or poor?" **(T2)** Bias Exposure Ablation: The phrase "The visual quality of the first image is poor" is removed from our prompt to examine the role of bias exposure, leading to the second prompt: "How about the visual quality of the second image? Good or poor?" **(T3)** Bias Mitigation Propagation Ablation: We delete the "How about" from the prompt to assess the impact of bias mitigation propagation to the query image, resulting in the third prompt: "The visual quality of the first image is poor. Rate the visual quality of the second image. Good or poor?" As shown in Table 4, the results reveal that: 1) Without our instructive prompt, even with the conditional images provided, bias cannot be effectively mitigated. 2) Without the explicit indication that the conditional images are of poor quality, the LMM fails to recognize the bias exposure, resulting in a marked performance drop. 3) The phrase "How about" suggests that the LMM should infer the query image's quality based on the prior understanding that the conditional images are of poor quality. Without this phrase, the propagation of bias mitigation from the conditional images to the query image weakens noticeably. The best results are achieved when the full prompt is included, demonstrating the necessity of each instruction in our prompt.

Study of Aggregation Scheme. To aggregate the quality scores derived from different conditional images, we introduce a semantic similarity aggregation strategy. To assess its effectiveness, we compare our method with four alternative schemes: **1) Average Aggregation:** Each of the four quality scores is assigned an equal weight during aggregation. **2) Quality Similarity Aggregation:** We utilize the widely adopted FR-IQA model, LPIPS (Zhang et al. 2018), to measure the quality similarity between the query image and each of its conditional images. These quality similarity scores are then treated as weights for the aggregation. **3) Winner-Takes-All:** The final quality score is determined solely by the quality score obtained from the conditional image that exhibits the highest semantic similarity to the query image. The results reveal that: 1) The Average scheme results in a noticeable performance drop (SRCC/PLCC 0.789/0.754 on LIVE Challenge, 0.824/0.838 on KonIQ-10k and 0.711/0.691 on AGIQA-3k), suggesting that uniform weighting fails to account for the bias variations across different types of distortions. 2) The Quality Similarity scheme is also ineffective (SRCC/PLCC 0.785/0.740 on LIVE Challenge, 0.817/0.827 on KonIQ-10k and 0.710/0.683 on AGIQA-3k). The possible reason may lie that a higher quality similarity score does not always correspond to a higher semantic recognition for the LMM, due to perceptual discrepancy between the LMM and the human visual system. 3) The Winner-Takes-All scheme, though commonly used for score aggregation, demonstrates suboptimal performance (SRCC/PLCC 0.632/0.491 on LIVE Challenge, 0.750/0.660 on KonIQ-10k and 0.623/0.529 on AGIQA-3k) as it fails to adequately capture the nuanced contributions of different conditional images. In comparison, our semantic similarity aggregation scheme delivers the best performance across, demonstrating superior generalization on diverse image distortions.

Conclusion

In this paper, we propose a training-free scheme to enhance the LMM in the IQA task. In particular, the perception bias that the LMM infers image quality highly relies on image semantics is mitigated by introducing conditional images in the prompt. These conditional images share the similar semantics as the query image but experience degraded quality. By instructing the LMM to align its quality ratings on those conditional images, the alignment in turn forces the LMM to rectify their judgment on the query image. Experimental results on images with different distortions verify the effectiveness of our method, and the generalization capability of our scheme across other LMMs highlights the potential for advanced prompt designs to fully leverage LMM knowledge for unseen tasks.

Acknowledgments

The work was supported by the National Natural Science Foundation of China under Grant No. 62401214.

References

Agnolucci, L.; Galteri, L.; Bertini, M.; and Del Bimbo, A. 2024. ARNIQA: Learning distortion manifold for image quality assessment. In *IEEE Winter Conference on Applications of Computer Vision*, 189–198.

Babu, N. C.; Kannan, V.; and Soundararajan, R. 2023. No reference opinion unaware quality assessment of authentically distorted images. In *IEEE Winter Conference on Applications of Computer Vision*, 2459–2468.

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Chen, B.; Xiao, K.; Shen, X.; and Wang, S. 2025. Monotonic and Invertible Network: A General Framework for Learning IQA Model from Mixed Datasets. *International Journal of Computer Vision*, 1–22.

Chen, B.; Zhu, L.; Li, G.; Lu, F.; Fan, H.; and Wang, S. 2021. Learning generalized spatial-temporal deep feature representation for no-reference video quality assessment. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(4): 1903–1916.

Chen, C.; Yang, S.; Wu, H.; Liao, L.; Zhang, Z.; Wang, A.; Sun, W.; Yan, Q.; and Lin, W. 2024. Q-Ground: Image Quality Grounding with Large Multi-modality Models. In *ACM International Conference on Multimedia*, 486–495.

Fang, Y.; Zhu, H.; Zeng, Y.; Ma, K.; and Wang, Z. 2020. Perceptual Quality Assessment of Smartphone Photography. In *IEEE Conference on Computer Vision and Pattern Recognition*, 3677–3686.

Fournier, A.; Fussell, D.; and Carpenter, L. 1982. Computer rendering of stochastic models. *Communications of the ACM*, 25(6): 371–384.

Ghadiyaram, D.; and Bovik, A. C. 2015. Massive online crowdsourced study of subjective and objective picture quality. *IEEE Transactions on Image Processing*, 25(1): 372–387.

Golestaneh, S. A.; Dadsetan, S.; and Kitani, K. M. 2022. No-Reference Image Quality Assessment via Transformers, Relative Ranking, and Self-Consistency. In *IEEE Winter Conference on Applications of Computer Vision*, 3209–3218.

Hendrycks, D.; and Dietterich, T. 2019. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. *International Conference on Learning Representations*.

Hosu, V.; Lin, H.; Sziranyi, T.; and Saupe, D. 2020. KonIQ-10k: An Ecologically Valid Database for Deep Learning of Blind Image Quality Assessment. *IEEE Transactions on Image Processing*, 29: 4041–4056.

Ke, J.; Wang, Q.; Wang, Y.; Milanfar, P.; and Yang, F. 2021. MUSIQ: Multi-scale image quality transformer. In *IEEE international conference on computer vision*, 5148–5157.

Li, B.; Zhang, Y.; Guo, D.; Zhang, R.; Li, F.; Zhang, H.; Zhang, K.; Li, Y.; Liu, Z.; and Li, C. 2024a. LLaVA-OneVision: Easy Visual Task Transfer. arXiv:2408.03326.

Li, C.; Zhang, Z.; Wu, H.; Sun, W.; Min, X.; Liu, X.; Zhai, G.; and Lin, W. 2023. AGIQA-3K: An Open Database for AI-Generated Image Quality Assessment. *IEEE Transactions on Circuits and Systems for Video Technology*, 1–1.

Li, F.; Zhang, R.; Zhang, H.; Zhang, Y.; Li, B.; Li, W.; Ma, Z.; and Li, C. 2024b. LLaVA-NeXT-Interleave: Tackling Multi-image, Video, and 3D in Large Multimodal Models. arXiv:2407.07895.

Lin, H.; Hosu, V.; and Saupe, D. 2019. KADID-10k: A Large-scale Artificially Distorted IQA Database. In *International Conference on Quality of Multimedia Experience*, 1–3.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024. Improved baselines with visual instruction tuning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 26296–26306.

Liu, Y.; Gu, K.; Li, X.; and Zhang, Y. 2020. Blind image quality assessment by natural scene statistics and perceptual characteristics. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 16(3): 1–91.

Luo, Y.; Yang, Z.; Meng, F.; Li, Y.; Zhou, J.; and Zhang, Y. 2023. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *arXiv preprint arXiv:2308.08747*.

Mittal, A.; Moorthy, A. K.; and Bovik, A. C. 2012. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12): 4695–4708.

Mittal, A.; Soundararajan, R.; and Bovik, A. C. 2012. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3): 209–212.

Moorthy, A. K.; and Bovik, A. C. 2010. A two-step framework for constructing blind image quality indices. *IEEE Signal Processing Letters*, 17(5): 513–516.

Ni, Z.; Liu, Y.; Ding, K.; Yang, W.; Wang, H.; and Wang, S. 2024. Opinion-Unaware Blind Image Quality Assessment using Multi-Scale Deep Feature Statistics. *IEEE Transactions on Multimedia*.

OpenAI. 2023. OpenAI. GPT-4V(ision) system card. Technical report, OpenAI.

Shang, C.; You, A.; Subramanian, S.; Darrell, T.; and Herzig, R. 2024. TraveLER: A Modular Multi-LMM Agent Framework for Video Question-Answering. *arXiv preprint arXiv:2404.01476*.

Sun, Y.; Zhang, Z.; Wu, H.; Liu, X.; Lin, W.; Zhai, G.; and Min, X. 2024. Explore the Hallucination on Low-level Perception for MLLMs. *arXiv preprint arXiv:2409.09748*.

Wang, J.; Chan, K. C.; and Loy, C. C. 2023. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 2555–2563.

Wang, X.; Chen, G.; Qian, G.; Gao, P.; Wei, X.-Y.; Wang, Y.; Tian, Y.; and Gao, W. 2023. Large-scale multi-modal pre-trained models: A comprehensive survey. *Machine Intelligence Research*, 20(4): 447–482.

Wang, Z.; and Bovik, A. C. 2006. *Modern image quality assessment*. Ph.D. thesis, Springer.

Wu, H.; Zhang, Z.; Zhang, E.; Chen, C.; Liao, L.; Wang, A.; Li, C.; Sun, W.; Yan, Q.; Zhai, G.; et al. 2023a. Q-bench: A benchmark for general-purpose foundation models on low-level vision. *arXiv preprint arXiv:2309.14181*.

Wu, H.; Zhang, Z.; Zhang, W.; Chen, C.; Liao, L.; Li, C.; Gao, Y.; Wang, A.; Zhang, E.; Sun, W.; et al. 2023b. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*.

Wu, Z.; Chen, X.; Pan, Z.; Liu, X.; Liu, W.; Dai, D.; Gao, H.; Ma, Y.; Wu, C.; Wang, B.; Xie, Z.; Wu, Y.; Hu, K.; Wang, J.; Sun, Y.; Li, Y.; Piao, Y.; Guan, K.; Liu, A.; Xie, X.; You, Y.; Dong, K.; Yu, X.; Zhang, H.; Zhao, L.; Wang, Y.; and Ruan, C. 2024. DeepSeek-VL2: Mixture-of-Experts Vision-Language Models for Advanced Multimodal Understanding. arXiv:2412.10302.

Ye, J.; Xu, H.; Liu, H.; Hu, A.; Yan, M.; Qian, Q.; Zhang, J.; Huang, F.; and Zhou, J. 2024. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. *arXiv preprint arXiv:2408.04840*.

Yu, W.; Yang, Z.; Li, L.; Wang, J.; Lin, K.; Liu, Z.; Wang, X.; and Wang, L. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conference on Computer Vision and Pattern Recognition*, 586–595.

Zhu, H.; Sui, X.; Chen, B.; Liu, X.; Chen, P.; Fang, Y.; and Wang, S. 2024. 2AFC prompting of large multimodal models for image quality assessment. *arXiv preprint arXiv:2402.01162*.